CRYSTAL GROWTH & DESIGN

Article

Prediction of Solid State Properties of Co-crystals using Artificial Neural Network Modelling

Rama Krishna Gamidi, Marko Ukrainczyk, Jacek Zeglinski, and Åke C. Rasmuson

Cryst. Growth Des., **Just Accepted Manuscript** • DOI: 10.1021/acs.cgd.7b00966 • Publication Date (Web): 07 Nov 2017

ACS Publications

Crystal Growth & Design is published by the American Chemical Society. 1155 Sixteenth Street N.W., Washington, DC 20036

Prediction of Solid State Properties of Co-crystals using Artificial Neural Network Modelling

Gamidi Rama Krishna, Marko Ukrainczyk, Jacek Zeglinski and Åke. C. Rasmuson*

Department of Chemical and Environmental Science, Synthesis and Solid State Pharmaceutical Centre, Bernal Institute, University of Limerick, Limerick, Ireland

Keywords. Co-crystals, Melting point, Lattice energy, Crystal density, Prediction, ANN models

Abstract. Using Artificial Neural Networks (ANNs), four distinct models have been developed for the prediction of solid-state properties of cocrystals: melting point, lattice energy, and crystal density. The models use three input parameters for the pure model compound (MC) and three for the pure coformer. In addition, as input parameter the model uses the p*K*a difference between the MC and the coformer, and a 1:1 MC–conformer binding energy as calculated by a force field method. Notably the models require no data for the actual cocrystals. In total, 61 CCs (two-component molecular cocrystals) were used to construct the models, and melting temperatures and crystal densities were extracted from the literature for four MCs: caffeine, theophylline, nicotinamide and isonicotinamide. The data set includes 14 caffein cocrystals, 9 theophylline cocrystals, 9 nicotinamide cocrystals and 29 isonicotinamide cocrystals. The model–I is trained using known cocrystal melting temperatures, lattice energies and crystal densities, to predict all three solid–state properties simultaneously. The average relative deviation for the training set is 2.49%, 6.21% and 1.88% for the melting temperature, lattice energy and crystal density, respectively, and correspondingly 6.26%, 4.58% and 0.99% for the valdation set. Model–II, model–III and model–IV were built using the same input neurons as in model–I, for separate prediction of each respective output solid–state property. For these models the average relative deviation for the traning sets becomes 1.93% for the melting temperature model-II, 1.29% for the lattice energy model-III and 1.03% for the crystal density model-IV, and correspondingly 2.23%, 2.40% and 1.77% for the respective validation sets.

Introduction

In the early stage of the drug discovery and development, the melting point (*T*m) of a compound is considered to be the first and most reliable physical property[1], *T*m is useful to estimate other properties[2, 3] such as vapor pressure,[4] boiling point,[5] intrinsic solubility[1, 2] and consequently bioavilability,[6] *etc*. Chu et al.[6] found a correlation between *T*m and the amount of dose absorbed of poorly soluble drugs of BCS class II and class IV systems–the lower the *T*m the more likely the drug will be well absorbed, and the less likely it is to face severe problems with bioavailability.[6] Moreover, consideration of *T*m of the substance is important in the pharmaceutical industry in order to set the processing parameters like handling, storage and disposal. Over the years, attempts have been made to estimate *T*m of new solid substances prior to the synthesis, e.g. by Quantitative Structure-Property Relationships (QSAR)[7, 8, 9, 10] and by the commercially available software programs based on different molecular descriptors.[11] However, better results were reported for structurally related components, *i.e*. homologous series of components rather than non-homologous series of components.[12, 13, 14, 15] Hence, those methods are yet not attractive for potential practical applications, particularly not for non-homologous series of components.

There is a number of reports on cocrystals (CC), especially, for the purpose of improving the physico-chemical properties[16, 17, 18] of a drug without modifying the drug molecule itself. In a cocrystal the modification occurs at the supramolecular level *via* intermolecular hydrogen bonding in the crystal lattice.[19] However, it has been concluded that the properties of synthesized pharmaceutical

CCs depend upon the judicious selection of the coformers. In example, the *T*m of pharmaceutical CCs can be controlled in a systematic manner by co-crystallization with a series of structurally related coformers[13, 20]. If one wishes to improve the thermal stability of a given Active Pharmaceutical Ingredient (API), then a coformer with higher *T*m is used and vice versa. However, the explanations are essentially qualitative rather than quantitative. Very little has been reported in the literature to quantify the physico-chemical properties of the CCs with respect to various coformers.[21] Because CCs are more complex systems than single component molecular systems, prediction of properties with respect to various coformers become even more challenging.

Estimation of *T*m of CCs prior to the synthesis could save cost and time, and help to screen libraries of new solid materials within the target range. In our previous work[22], we reported an ANN QSAR model for estimation of *T*m of the CCs with a good correlation capability. However, this model is a correlation model since among the input variables we use data that can only be acquired by experimental measurements on the actual CCs. From application point of view, it is much more useful when a model can be used for actual prediction of the melting point, without requiring actual measured data for the CCs and thus without the CC actually being manufactured. Accordingly, in the present work, such a model for prediction of *T*m of CCs is intended and succeeded using ANN methods. In addition to the *T*m, we also succeed to predict two more solid –state properties of the CCs, the lattice energy and crystal density. Prediction of the lattice energy is a first step towards prediction of the melting enthalpy and thus of the solid phase free energy of fusion (and eventually the solid-liquid solubility) of the CCs. Concurrently, E_{latt} of the CCs can be used to examine stabilization or destabilization of the solid phase *via* CCs formation compared to the pure API solid. Crystal density play a role in comparision of many of the physical properties of a substances such as stability (more stable form would expect to have higher CD, especially in case of polymorphs) and melting point. Higher CD depicts the existence of close molecular arrangement through in–combination of $\pi\cdots\pi$ stacking and stronger intermolecular interactions, which corresponds to higher stability thereby higher melting point. In this work, 61 CCs of four different MC's caffeine (CAF), theophylline (THP), isonicotinamide (INA) and nicotinamide (NA) (14-CAF, 9-THP, 29-INA and 9-NA) were selected (Figure 1). The rationality for selecting these four molecules is that, CAF, THP and NA molecules are APIs, while INA regarded as a GRAS (Generally Recognized As Safe) coformer. Moreover, all four components have plenty of cocrystal reports available in the literature. The information on individual components and respective CCs were retrieved from the literature using the Scifinder and Cambridge Structural Database software. Lattice and binding energies were calculated for all selected CCs and individual components by using the COMPASS II forcefield. Four different ANN models have been built to predict the three different solid-state properties of the CCs prior to the synthesis. In addition, sensitivity analysis with respect to each input neuron in the input layer of the all four models has been performed.

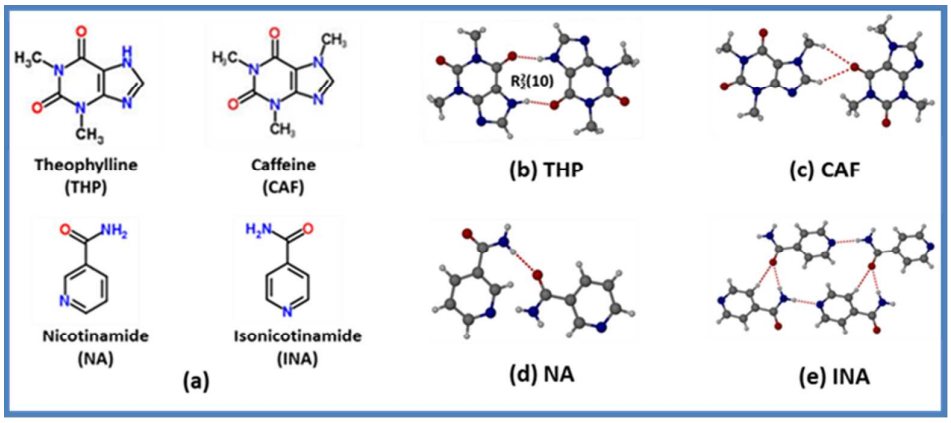


Figure 1. (a) Chemical structures of all four MCs, and the major synthon which is present in the stable solid form of respective MC, at ambient conditions were given as (b), (c), (d) & (e).

Methods and Calculations

Artificial Neural Network Modelling

Herein, Artificial Neural Network models[23] are used to predict the solid-state properties[24, 25, 26] of the cocrystals. The architecture of the constructed ANN model[24] is composed of an input layer, a hidden layer(s), weights, a sum function, an activation function and an output layer, as is illustrated in the Figure 2. A multiple–layer feed forward-back propagation network was used to flow the information in one direction (*i.e.* from input to output) and uses linear/nonlinear approximation functions effectively until it reaches to convergence criterion to make a relationship between inputs and output vectors, where back-propagation of error algorithm is employed to calculate ANN weights. The Gradient Descent method is applied to adjust the weight parameters to minimize the mean squared error between the experimental and the ANN predicted output solid–state properties of the network during the back propagation process. In addition, a logistic function and a purelin function were used as the propagation functions in the hidden layer and in the output layer, respectively. All input vectors and the output vectors were normalized before performing the training process, such that they fall in the interval range of 0–1, hence, their standard deviations will also fall within the range of value one.[24, 25, 26]

Neural network models are sensitive to the number of neurons in the hidden layer. A better fitting of the traning set will be obtained by using a higher number of neurons, but a higher number can lead to overfitting, which leads to larger deviation between the experimental and the predicted solid–state for the validation data set. To overcome this problem, the ANN predictive model was trained with one hidden layer, starting from using one neuron and gradually increasing the number. In each step the output values for both the training set and validation set were examined. By systematic evaluation, it was concluded that 8 neurons are sufficient for the hidden layer when the input layer contains eight neurons. The performance of the model does not increase much beyond eight neurons in the hidden layer, and accordingly the training process has fallen into the global minimum. Since the Kolmogotov theorem[27] states that less than two hidden layers are sufficient to build a model for any problem, and a higher number leads to over-fitting and poor generalization capability of the model. Therefore, the size of the constructed neural network for model–I is 8–8–3, whereas for model–II, model–III and model–IV is 8–8–1.

The constructed model–I is aimed to simultaneously predict the three different solid–state properties of the CCs: the melting temperature (*T*m), crystal density (CC_{CD}) and lattice energy (CC_{Elatt}) as output parameters using eight input parameters. Models II, III and IV were built to predict the *T*m (model–II), CC_{CD} (model–III) and CC_{Elatt} (model–IV), respectively to examine the efficiency of model–I. Among the eight operating variables the model uses data for the pure MC and coformer: the molecular weight (MC_{MW}), (CF_{MW}); the melting temperature (MC_{Tm}), (CF_{Tm}); and a pure compound binding energy (MC_{BE}), (CF_{BE}) as explained below in the binding energy calculations. In addition, the model uses the difference in p*K*a between the MC and the coformer, (Δp*K*a) as an input parameter: Δp*K*a = p*K*a (base) – p*K*a (acid). For complexes involving two acids, the p*K*a of the more basic compound (with more basic substituent's) is taken as p*K*a (base)). As the eigth input parameter, the model uses a MC–coformer binding energy. This binding energy is calculated by molecular simulation force field calculation over the binding of a 1:1 heterodimer in gas phase between the model molecule and the coformer molecule. The ANN model is schematically presented in Figure 2 and is developed using 61 CCs of four different MC's, and were divided into two sets: i) 55 data points for the training set, and ii) 6 data points for the validation set (as new data points for the prediction) containing one system from each of THP, CAF and NA (Saliylic acid (SA), 4-Fluoro-3-nitro aniline (4F3NAN) and Glutaric acid (GTA) respectively); three from INA (Adipic acid (ADP), 4-hydroxybenzoic acid (4HBA) and Glutaric acid (GTA). The training set is used to train the network, whereas the validation set is used to test the generalization of the model. It is noteworthy to mentioned that, ANN model is not constructed based on this valdation set, but is used to verify the strength of the model during the training process to avoid the overfitting of the model.

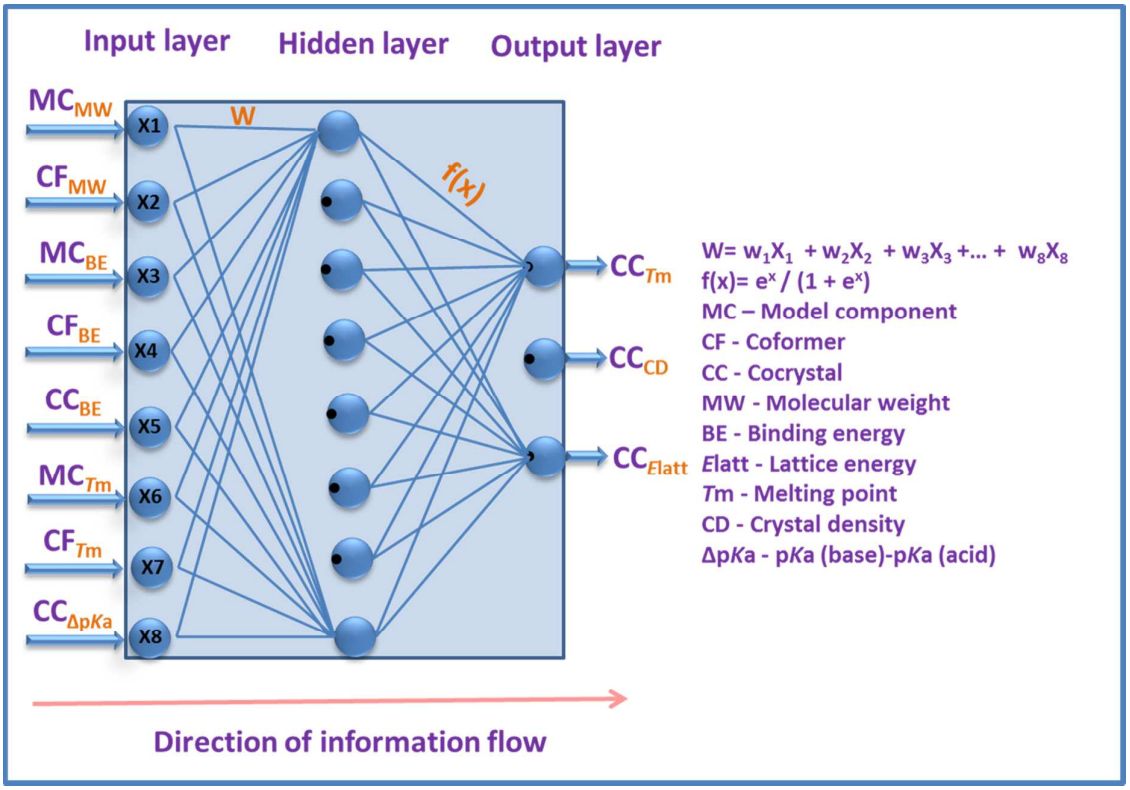


Figure 2. The architecture of the constructed ANN model, consisting of three main layers, input, hidden and output layer. The input layer is used to introduce the input variables to the network and output layer represents predictions of the response (output) variables calculated by ANN.

The training process

In the training process of the models, weight parameters were adjusted iteratively to minimize the criterion function. The attributes which are present in the input/output vectors were normalized between 0 and 1 (within the limitations of the sigmoid transfer function *i.e.*, logsig). The neurons present in the input layer (eight parameters) fed–in through connections with some random weights used from 0 to 1, and also the values used as for the learning parameters and momentum for generalization of all constructed four models are given in the Table 1. Herein, the total weight of the input layer is nothing but the weighted sum of all eight input parameters. Each neuron in the input layer is connected to all eight neurons in the hidden layer, thereafter, the information will transfer (through logistic transfer/activation function) into the output layer which for model-I contains three distinct solid– state properties of the CCs, *i.e. T*m, CC_{CD} and CC_{Elatt}, while for models II, III and IV it contains just each respective targeted property. To build and train the ANNs model, a Neural network software package[28] was employed.

Model	Weights	Learning parameter	Momentum
I	0.8	0.4	0.6
II	0.8	0.4	0.6
III	0.7	0.3	0.6
IV	0.8	0.3	0.8

Table 1. Weight parameters, learning parameters and momentum values used to build the four models.

Binding& Lattice energy calculations

Herein, the binding energy (BE)[29, 30] stands for the interaction energy between molecules forming a synthon. In crystal engineering Desiraju defined supramolecular synthon[31, 32, 33] as "structural units within supramolecules which can be formed and/or assembled by known or conceivable synthetic operations involving intermolecular interactions." Accordingly, a forcefield method is employed to calculate the BE for all MCs and pure coformers to be used as input data and for all 61 CCs. All the synthon dimers or clusters featuring the strong intermolecular H–bonds (such as X–H···Y–Z: X=O, N; Y=O, N and Z=C, H) were extracted from their respective crystal structures (the CSD Refcodes are available in the Table 1) and used as starting structures for the calculations (a list of representative supramolecular synthons listed in the SI Table 1). The synthons considered here for calculation were major synthons (that is providing main driving force for the formation of stable cocrystals), existing in their stable solid state configuration. The selected synthons were subsequently optimized in gas–phase using the COMPASS II forcefield as implemented in the Forcite module of the Material Studio software (Accerlys Inc.), and energies were calculated in fully relaxed gas–phase geometries. Thereafter, the BE for CCs (containing two or more molecules) is calculated according to equation 1.

$$\Delta E_{bind} = E_{A\text{-}B\ldots\text{-}N} - (E_A + E_B + \ldots + E_N) \text{ - } (1)$$

where $E_{A\text{-}B\ldots\text{-}N}$ is the energy of a synthon and E_A, E_B, and E_N are the energies of isolated monomers of A, B and N^{th} molecule

In some cases, after a full optimization cycle's, the initial geometry of a synthon changed from its in–plane orientation (favorable in the crystal packing arrangement) to more bent, *i.e.* out–of–

plane geometry. Such a deformation of a synthon implies significant change in its linearity, thus, it weakens the strength of the H–bonds and ultimately yields an energy value which is not relevant to the BE of a synthon existing in its crystal lattice. Therefore, in some cases the number of iteration cycles for the geometry optimization of a synthon was restricted to retain the cluster geometry as close as possible to its crystal–like in–plane molecular orientation. E_{latt} values for all 61 CCs were extracted from our previous work, wherein, calculations were performed (Table 2) by using the COMPASS II forcefield, as explained elesewhere.[22]

Database Creation

A database over CCs of four MCs, namely, CAF, THP, NA and INA have been extracted from our previous work, the method employed for the creation of database using the Scifinder and the Cambridge Structural Database softwares (CSD version 5.37, update 1 (Nov 2015) was explained eleswhere.[22]

Table 2. The 61CCs used in this study, respective CSD refcodes, stoichiometric ratio, Δp*K*a and E_{latt}.

Name of the Component	Code	p*K*a	Cocrystal	Δp*K*a	E_{latt} of CCs (Kcal/mol)	Ratio	Cocrystal Refcode
Caffeine	CAF	0.7 (cb)	-	-	-	-	-
Theophylline	THP	1.7 (cb) 8.77 (ca)	-	-	-	-	-
Isonicotinamide	INA	3.61 10.61	-	-	-	-	-
Nicotinamide	NA	3.35	-	-	-	-	-
DL-Malic acid	DLMA	3.40 5.11	THP:DLMA	-1.7 -3.66	-67.876	1:1	CIZTAH
D-Malic acid	DMA	3.40 5.11	THP:DMA	-1.7 -3.66	-67.197	1:1	CODCOO
Glutaric acid	GTA	4.31 5.41	THP:GTA	-2.6 -3.36	-60.167	1:1	XEJXIU
Gentisic acid	GNA	2.97	THP:GNA	-1.27	-63.764	1:1	DUCROJ
Salicylic acid	SA	2.97 13.82	THP:SA	-1.27 -12.12	-55.419	1:1	KIGLES
p-coumaric acid-I	PCA-I	4 9.51 *M	THP:PCA-I	-2.3 -7.81	-64.651	1:1	IJIBEJ
p-coumaric acid-II	PCA-II	"	THP:PCA-II	"	-63.937	1:1	IJIBEJ01
Saccharin	SAC	11.68	THP:SAC	-9.98	-59.602	1:1	XOBCUN
Urea	URE	0.10	THP:URE	1.60	-52.728	1:1	DUXZAX
Glutaric acid	GTA-I	4.31 5.41	CAF:GTA-I	-3.61 -4.71	-59.068	1:1	EXUQUJ01

Glutaric acid	GTA-II	"	CAF:GTA-II	"	-59.512	1:1	EXUQUJ
p-coumaric acid	PCA	4 9.51 *M	CAF:PCA	-2.3 -7.81	-63.746	1:1	IJEZUT
4-nitroaniline	4NAN	1	CAF:4NAN	-0.3	-53.417	1:1	LATGUK
2-iodo-4-nitroaniline	2I4NAN	0.46 *M	CAF:2I4NAN	0.24	-49.364	1:1	LATFUJ
2-fluoro-5-nitroaniline	2F5NAN	0.52 *M	CAF:2F5NAN	0.18	-52.866	1:1	LATHEV
4-fluoro-3-nitroaniline	4F3NAN	1.42 *M	CAF:4F3NAN	-0.72	-52.236	1:1	LATGIY
4-chloro-3-nitroaniline	4C3NAN	1.90	CAF:4C3NAN	-1.2	-57.170	1:1	LATGEU
2-chloro-5-nitroaniline	2C5NAN	0.40 *M	CAF:2C5NAN	0.3	-53.369	1:1	LATGOE
4-iodo-3-nitroaniline	4I3NAN	1.28 *M	CAF:4I3NAN	-0.58	-55.371	1:1	LATGAQ
2,4-dinitrobenzoic acid	24DNBA	1.43	CAF:24DNBA	-0.73	-61.841	1:1	LATHAR
2-fluoro-5-nitrobenzoic acid	2F5NBA	2.69 *M	CAF:2F5NBA	-1.99	-57.534	1:1	LATHIZ
Salicylic acid	SA	2.98 13.82	CAF:SA	-2.28 -13.12	-54.979	1:1	XOBCAT
Salicylic acid_I	SA-I	"	CAF:SA-I	"	-54.990	1:1	XOBCAT01
Oxalic acid	OXA	1.23 4.19	INA:OXA	2.38 -0.58	-84.971	2:1	ULAWAF
Malonic acid	MLA	2.83 5.69	INA:MLA	0.78 -2.08	-126.336	2:1	ULAWEJ
Succinic acid	SCA	4.16 5.61	INA:SCA	-0.55 -2	-85.669	2:1	LUNNUD
Glutaric acid	GTA	4.31 5.41	INA:GTA	-0.7 -1.8	-56.894	1:1	ULAXAG
Adipic acid	ADA	4.43 5.41	INA:ADA	-0.82 -1.8	-57.713	1:1	ULAXEK
Pimelic acid	PIA	4.71 5.58	INA:PIA	-1.1 -1.97	-58.609	1:1	ISIJEA
Suberic acid	SUA	4.52 5.49	INA:SUA	-0.91 -1.88	-62.187	1:1	ISIJIE
Azelaic acid	AZA	4.550 5.498	INA:AZA	-0.94 -1.88	-61.805	1:1	ISIJAW
Fumaric acid	FUA	3.03 4.44	INA:FUA	0.58 -0.83	-84.215	2:1	LUNNOX
4-ketopimelic acid	4KPIA	3.68 *M 4.42 *M	INA:4KPA	-0.07 -0.81	-91.711	2:1	LUNNIR

12-bromododecanoic acid	12BDA	4.95 *M	INA:12BDA	-1.34	-62.859	1:1	LUNMUC
Salicylic acid	SA	2.98 13.82	INA:SA	0.63 -10.21	-50.890	1:1	XAQQEM
3-hydroxybenzoic acid	3HBA	4.06 9.92	INA:3HBA	-0.45 -6.31	-53.994	1:1	LUNMEM
4-hydroxybenzoic acid	4HBA	4.48 9.32	INA:4HBA	-0.87 -5.71	-55.167	1:1	VAKTOR
4-fluorobenzoic acid	4FBA	4.15	INA:4FBA	-0.54	-49.167	1:1	ASAXUN01
3-nitrobenzoic acid	3NBA	3.47	INA:3NBA	0.14	-54.057	1:1	ASAXOH
2-hexenoic acid	2HEA	5.13 *M	INA:2HEA	-1.52	-48.183	1:1	AJAKAX
Cinnamic acid	CIA	3.89 (cis) 4.44 (trans)	INA:CIA	-0.28 -0.83	-52.787	1:1	LUNMAI
Chloroacetic acid	CAA	2.85	INA:CAA	0.76	-44.004	1:1	LUNNAJ
(RS)-2-phenylpropionic acid	2PPARS	4.34	INA:2PPARS	-0.73	-48.773	1:1	ROLFOO
(R)-2-phenylpropionic acid	2PPAR	4.34	INA:2PPAR	-0.73	-48.469	1:1	RONDAA
dl-mandelic acid	DLMDA	3.85	INA:DLMDA	-0.24	-56.255	1:1	LUNPAL
Clofibric acid	CFA	3.0	INA:CFA	0.61	-54.279	1:1	UMUYUX
Resorcinol	REOL	9.32 11.1	INA:REOL	-5.71 -7.49	-77.890	2:1	VAKTUX
Hydroquinone	HQ	9.85 11.4	INA:HQ	-6.24 -7.79	-76.792	2:1	VAKVIN
3-(N,N-dimethylamino)benzoic acid	3NNDMABA	3.76 *M 4.92 *M	INA: 3NNDMABA	-0.15 -1.31	-51.969	1:1	LUNMIQ
3,5-bis(trifluoromethyl)benzoic acid	35TFMBA	3.81 *M	INA:35TFMBA	-0.2	-51.324	1:1	LUNMOW
Meclofenamic acid	MEFA	3.79	INA:MEFA	-0.18	-62.583	1:1	SAXPAK
Fumaric acid monoethyl ester	FAMEE	3.48 *M	INA:FAMEE	0.13	-51.755	1:1	LUNNEN
Fumaric acid	FUA	3.03 4.44	NA:FUA	0.32 -1.09	-54.457	1:1	NUKYAU
Glutaric acid	GTA	4.31 5.41	NA:GTA	-0.96 -2.06	-57.708	1:1	NUKYEY
4-hydroxybenzoic acid	4HBAII	4.48 9.32	NA:4HBAII	-1.13 -5.97	-53.731	1:1	RUYHEZ01
Ethyl paraben	EPB	8.34	NA:EPB	-4.99	-51.664	1:1	GOGQID
2-chloro-4-nitrobenzoic acid	2C4NBA	0.94	NA:2C4NBA	2.41	-54.546	1:1	SUTTUX
Tolfenamic acid	TOFA	3.88	NA:TOFA	-0.53	-87.578	2:1	EXAQIE

Mefenamic acid	MEFA	3.79	NA:MEFA	-0.44	-85.931	2:1	EXAQOK
Niflumic acid	NIFA	1.88	NA:NIFA	1.47	-62.731	1:1	EXAQEA
Furosemide	FURA	4.25	NA:FURA	-0.9	-76.015	1:1	YASGOQ

*M-Calculated using the Marvin Sketch software

Results and Discussion

The constructed predictive QSAR model–I consists of eight neurons in the input layer and three neurons in the output layer. One and the same model–I is trained and validated for predicting the three different solid–state properties of the CCs simultaneously. The selected eight input parameters were the most influential parameters on the outcome of the three solid–state properties of the CCs: melting point, lattice energy and crystal density. However, this conclusion was reached on the basis of trial and error. The initial training process started by considering the six parameters (MC_{MW}, CF_{MW}, functional group which is present in the MC (MC_{FG}), type of functional group which is present in the coformer (CF_{FG}), MC_{Tm} and CF_{Tm}) as input neurons in the input layer, which gave an average relative error (ARE) of 11.8% for the training set and 14.2% for the validation set. These values are averages of relative error over all 61 systems and all three output variables. By addition of Δp*K*a these ARE values were reduced to 7.86% error for the training set and 9.36% error for the validation set, and this result was better than any attempt to use Δp*K*a to replace one of the initial six parameters. However, to improve the model further, the synthon energy or binding energies of the MC_{BE}, CF_{BE} and CC_{BE} were included as three additional input neurons, and MC_{FG} and CF_{FG} were removed. Thus, this model with 8 input neurons improved the fit such that the training set error reduced to 3.53% and the validation set error reduced to 3.95%. Inclusion of also MC_{FG} and CF_{FG} into the input layer (10 neurons) reduces the training set error to 2.61% but the validation set error increases to 7.75%, and thus these two parameters were deemed to not improve the overall performance. The high deviation for the validation set compared to the training set is seen as due to overfitting of the model.

The training process was stopped after reaching into the convergence criterion with 3.53% average relative error for the training set and 3.95% for the validation set, again values being aggregate relative deviations of model–I for prediction of the three solid–state properties of CCs simultaneously. In examining the contribution from each individual output parameter, the calculated relative deviations for the prediction of *T*m of the CCs is 2.49% for the training set and 6.26% for the validation set; for CC_{Elatt} the value is 6.21% for the training set and 4.58% for the validation set; and for prediction of CC_{CD} it is 1.88% for the training set and 0.99% for the validation set. This can be due to that the training set algorithm converges to a local minimum, that happens to be the global minimum for the validation set. For the lattice energy and the crystal density, the validation set deviation is lower than the training set deviation.The capability of the predictive ANN model–I towards the prediction of *T*m, E_{latt} and CD of the CCs is shown in the Figure (3 – 5), and values are given in Tables 3, 5 and 10.

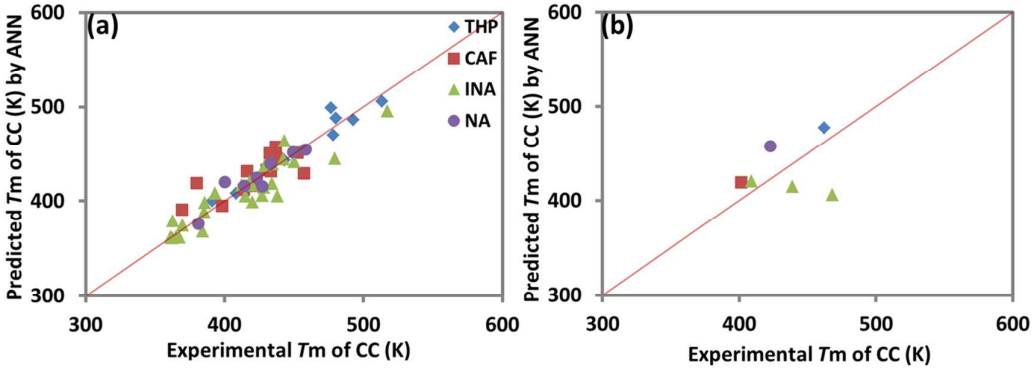


Figure 3. Model–I predicted *vs* experimental melting temperature (eigth input parameters): (a) training set, (b) validation set.

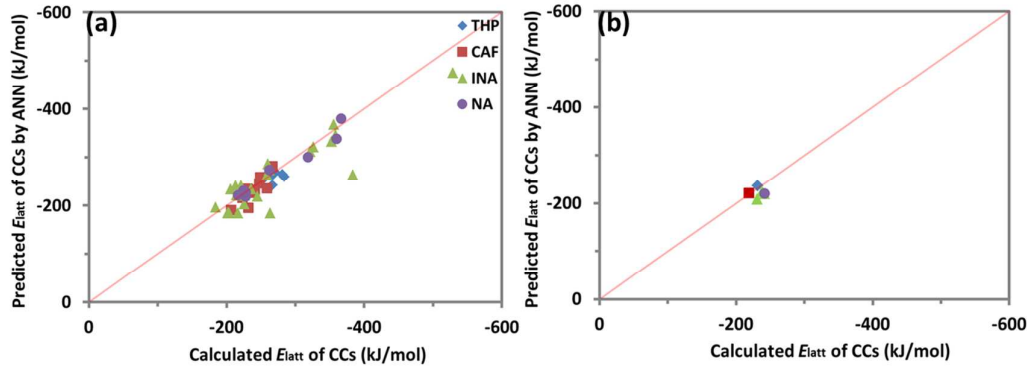


Figure 4. Model–I predicted *vs* experimental lattice energy (eigth input parameters): (a) training set, (b) validation set.

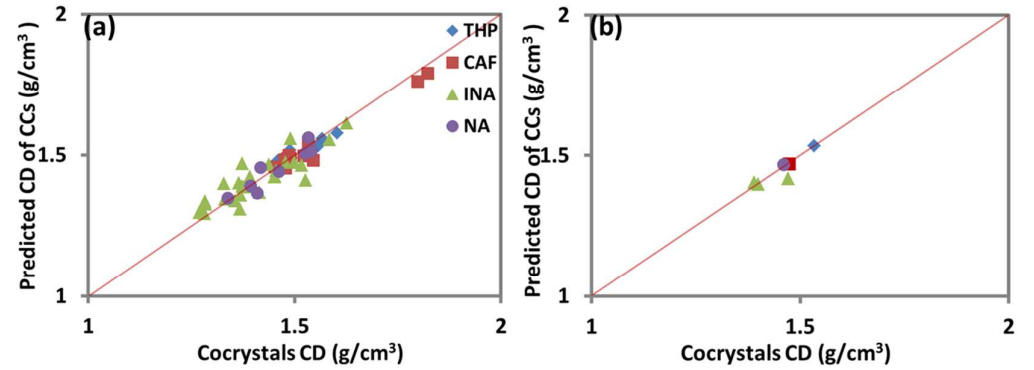


Figure 5. Model–I predicted *vs* experimental crystal density (eigth input parameters): (a) training set, (b) validation set.

Table 3. Experimental and predicted *T*m of the 61 CCs of CAF, THP, NA and INA drug molecules.

Name of the CC	MC (*T*m) (K)	CF (*T*m) (K)	*T*m of the CC (exp) (K)	*T*m of the CC (pre) (K), model-I	*T*m (pre)- *T*m (exp) (K), model-I	*T*m of the CC (pre) (K), model-II	*T*m (pre)- *T*m (exp) (K), model-II
Training set							
THP: DLMA	544.2	403.2	443.2	443.0	-0.2	439.8	-3.4
THP: DMA	544.2	371.7	**408.2**	**408.2**	**0.1**	405.9	-2.3
THP:GTA	544.2	369.7	391.2	399.8	8.6	402.6	11.4

THP:GNA	544.2	475.7	513.2	505.6	-7.5	**497.8**	**-15.4**
THP:PCA-I	544.2	484.7	492.8	486.2	-6.5	484.7	-8.1
THP:PCA-II	544.2	484.7	**476.8**	**499.1**	**22.4**	488.9	12.1
THP:SAC	544.2	502.0	480.2	488.1	7.9	481.8	1.6
THP:URE	544.2	406.2	478.2	470.4	-7.8	**478.1**	**-0.1**
CAF:GTA_I	509.7	369.7	398.2	394.7	-3.5	392.6	-5.4
CAF:GTA_II	509.7	369.7	369.2	390.8	21.6	390.5	21.4
CAF: PCA	509.7	484.7	452.6	451.1	-1.4	450.2	-2.3
CAF:4NAN	509.7	420.7	436.9	457.6	20.7	432.2	-4.6
CAF:2I4NAN	509.7	380.2	430.2	432.6	2.4	**429.8**	**-0.4**
CAF:2F5NAN	509.7	371.7	413.7	412.3	-1.5	400.7	-13.0
CAF:4C3NAN	509.7	373.2	420.7	416.2	-4.5	407.6	-13.1
CAF:2C5NAN	509.7	394.2	**379.7**	**418.7**	**39.1**	407.2	27.5
CAF:4I3NAN	509.7	415.2	438.2	444.6	6.5	440.0	1.8
CAF:24DNBA	509.7	454.2	432.4	450.6	18.2	449.7	17.3
CAF:2F5NBA	509.7	416.2	457.2	429.2	-27.9	**424.3**	**-32.9**
CAF:SA	509.7	432.2	416.0	431.5	15.5	430.3	14.3
CAF:SA_I	509.7	432.2	**433.2**	**431.8**	**-1.3**	430.5	-2.7
INA:OXA	429.2	375.7	517.0	495.6	-21.4	508.9	-8.1
INA:MLA	429.2	409.2	443.2	463.9	20.8	440.1	-3.1
INA:SCA	429.2	457.2	**479.2**	**444.5**	**-34.6**	**450.5**	**-28.7**
INA:PIA	429.2	377.2	385.2	388.6	3.5	397.3	12.1
INA:SUA	429.2	415.7	438.2	405.0	-33.2	409.7	-28.5
INA:AZA	429.2	382.2	415.2	404.9	-10.2	414.4	-0.8
INA:FUA	429.2	560.2	420.2	416.8	-3.4	427.1	6.9
INA:4KPA	429.2	416.2	385.7	398.0	12.3	386.0	0.3
INA:12BDA	429.2	326.7	362.2	361.1	-1.1	363.6	1.4
INA:SA	429.2	431.8	393.2	407.7	14.6	405.3	12.1
INA:3HBA	429.2	472.2	418.2	420.2	2.0	414.2	-4.0
INA:4FBA	429.2	457.2	427.2	405.5	-21.6	430.6	3.6
INA:3NBA	429.2	413.2	434.2	418.4	-15.7	431.7	-2.5
INA:2HEA	429.2	307.2	384.2	367.7	-16.4	376.3	-7.9
INA:CIA	429.2	406.2	420.2	398.7	-21.4	395.1	-24.9
INA:CAA	429.2	336.2	369.7	373.8	4.2	377.8	8.1
INA:2PPARS	429.2	302.7	365.0	361.2	-3.8	361.8	-3.2
INA:2PPAR	429.2	302.7	**361.0**	**361.1**	**0.1**	361.7	0.7
INA:DLMDA	429.2	403.2	442.2	444.1	2.1	445.1	2.9
INA:CFA	429.2	393.7	362.7	378.6	16.0	369.7	7.0
INA:REOL	429.2	383.2	428.2	414.6	-13.6	429.8	1.6
INA:HQ	429.2	445.2	429.0	437.3	8.3	420.3	-8.7
INA: 3NNDMABA	429.2	423.7	412.2	416.8	4.6	424.5	12.3
INA:35TFMBA	429.2	415.2	434.7	439.2	4.5	429.3	-5.4
INA:MEFA	429.2	522.2	450.0	441.2	-8.7	**450.1**	**0.1**
INA:FAMEE	429.2	337.7	367.5	361.6	-5.9	365.1	-2.4
NA:FUA	401.2	560.2	449.2	451.8	2.6	445.9	-3.3
NA:4HBAII	401.2	487.7	458.2	454.3	-3.9	463.7	5.6
NA:EPB	401.2	389.7	381.0	375.7	-5.3	383.6	2.6
NA:2C4NBA	401.2	412.7	432.8	439.6	6.9	430.0	-2.8
NA:TOFA	401.2	480.2	427.0	415.5	-11.5	422.3	-4.7
NA:MEFA	401.2	503.7	**400.0**	**419.8**	**19.8**	**413.4**	**13.4**
NA:NIFA	401.2	477.2	414.0	416.3	2.3	**415.9**	**1.9**
NA:FURA	401.2	493.2	**423.2**	**424.4**	**1.2**	419.6	-3.6
Validation set							
THP:SA	544.2	432.2	462.2	477.6	15.5	474.3	12.2
CAF:4F3NAN	509.7	368.2	401.7	419.3	17.7	**405.6**	**3.9**
INA:ADA	429.2	425.3	439.0	414.9	-24.1	432.6	-6.4
INA:GTA	429.2	369.7	**409.0**	**420.6**	**11.6**	**423.1**	**14.1**
INA:4HBA	429.2	487.7	**468.2**	**405.9**	**-62.3**	454.8	-13.4
NA:GTA	401.2	369.7	423.0	457.6	34.7	428.5	5.6

Table 4. A detailed analysis of prediction of *T*m, E_{latt} and CD of the 61 CCs using model–I, model–II, model–III and model–IV showing the lowest and highest deviation between model and experimental values.

Melting point								
Constructed model	**model–I**				**model–II**			
	lowest	Deviation %	highest	Deviation %	lowest	Deviation %	highest	Deviation %
training set	THP: DMA	0.1	CAF: 2C5NAN	39.1	THP:URE	-0.1	CAF: 2F5NBA	-32.9
	INA:2PPAR	0.1			INA: MEFA	0.1		
validation set	INA:GTA	11.6	INA: 4HBA	-62.3	CAF: 4F3NAN	3.9	INA:GTA	14.1
Lattice energy								
Constructed model	**model–I**				**model–III**			
	lowest	Deviation %	highest	Deviation %	lowest	Deviation %	highest	Deviation %
training set	THP:GTA	-0.1	INA: 4KPA	28.5	NA: 2C4NBA	0.0	INA:4KPA	33.3
validation set	CAF: 4F3NAN	-0.7	INA: 4HBA	-49.9	CAF: 4F3NAN	-0.6	INA:ADA	12.9
Crystal density								
Constructed model	**model–I**				**model–IV**			
	lowest	Deviation %	highest	Deviation %	lowest	Deviation %	highest	Deviation %
training set	CAF:2F5NAN	0.000	INA: 3NBA	-0.116	THP: PCAII	0.000	INA:3NBA	-0.119
			INA:CIA	-0.116	CAF: PCA	0.000		
validation set	THP:SA	0.000	INA:4HBA	-0.054	THP:SA	-0.001	NA:GTA	-0.048

Table 5. Experimental and predicted lattice energies of the 61 CCs of CAF, THP, NA and INA drug molecules.

Name of the CC	**MC$_{Elatt}$ (kJ/mol)**	**CC$_{Elatt}$ (Cal) (kJ/mol)**	**CC$_{Elatt}$ (Pre) (kJ/mol), model–I**	**CC$_{Elatt}$ (pre) - CC$_{Elatt}$ (exp) (kJ/mol), model–I**	**CC$_{Elatt}$ (Pre) (kJ/mol), model–III**	**CC$_{Elatt}$ (pre) - CC$_{Elatt}$ (exp) (kJ/mol), model–III**
Training set						
THP: DLMA	-140.2	-284.1	-259.8	24.3	**-269.0**	**15.1**
THP: DMA	"	-281.2	-265.3	15.9	-273.2	7.9
THP:GTA	"	**-251.9**	**-252.3**	**-0.4**	-260.2	-8.4
THP:GNA	"	**-266.9**	**-242.3**	**24.7**	-272.4	-5.4
THP:PCA-I	"	-270.7	-269.4	1.3	-268.2	2.5
THP:PCA-II	"	-267.4	-264.4	2.9	-265.3	2.1
THP:SAC	"	-249.4	-252.3	-2.9	**-250.6**	**-1.3**
THP:URE	"	-220.5	-224.7	-4.2	-225.5	-5.0
CAF:GTA_I	-128.0	**-247.3**	**-246.0**	**1.3**	**-246.9**	**0.4**
CAF:GTA_II	"	-248.9	-259.4	-10.5	**-261.5**	**-12.6**
CAF: PCA	"	-266.5	-282.0	-15.5	-270.3	-3.8
CAF:4NAN	"	-223.4	-215.9	7.5	-218.4	5.0
CAF:2I4NAN	"	-206.3	-190.4	15.9	-216.7	-10.5
CAF:2F5NAN	"	-221.3	-229.3	-7.9	-224.7	-3.3
CAF:4C3NAN	"	-239.3	-226.4	13.0	-227.6	11.7
CAF:2C5NAN	"	-223.4	-227.6	-4.2	-216.7	6.7
CAF:4I3NAN	"	**-231.8**	**-195.4**	**36.4**	-237.7	-5.9
CAF:24DNBA	"	-258.6	-235.6	23.0	-248.5	10.0

CAF:2F5NBA	"	-240.6	-232.6	7.9	-234.7	5.9
CAF:SA	"	-230.1	-234.7	-4.6	-233.9	-3.8
CAF:SA_I	"	-230.1	-235.1	-5.0	-235.1	-5.0
INA:OXA	-114.2	-355.6	-367.8	-12.1	-358.6	-2.9
INA:MLA	"	-528.4	-474.9	53.6	-489.9	38.5
INA:SCA	"	-358.6	-345.6	13.0	-357.3	1.3
INA:PIA	"	-245.2	-218.4	26.8	-243.1	2.1
INA:SUA	"	-260.2	-287.0	-26.8	-269.9	-9.6
INA:AZA	"	-258.6	-262.8	-4.2	-251.0	7.5
INA:FUA	"	-352.3	-332.2	20.1	-338.5	13.8
INA:4KPA	"	-383.7	**-264.4**	**119.2**	**-244.3**	**139.3**
INA:12BDA	"	-263.2	-184.1	79.1	-256.5	6.7
INA:SA	"	-213.0	-241.4	-28.5	-232.2	-19.2
INA:3HBA	"	-225.9	-202.9	23.0	-208.8	17.2
INA:4FBA	"	-205.9	-234.3	-28.5	-241.8	-36.0
INA:3NBA	"	-226.4	-220.5	5.9	-219.7	6.7
INA:2HEA	"	-201.7	-184.9	16.7	-187.4	14.2
INA:CIA	"	-220.9	-241.4	-20.5	-238.5	-17.6
INA:CAA	"	-184.1	-196.2	-12.1	-196.6	-12.6
INA:2PPARS	"	-204.2	-184.1	20.1	-184.1	20.1
INA:2PPAR	"	-202.9	-184.1	18.8	-184.1	18.8
INA:DLMDA	"	**-235.6**	**-234.3**	**1.3**	**-235.1**	**0.4**
INA:CFA	"	-227.2	-224.3	2.9	-239.7	-12.6
INA:REOL	"	-325.9	-321.7	4.2	-324.3	1.7
INA:HQ	"	-321.3	-311.7	9.6	-319.7	1.7
INA:3NNDMABA	"	-217.6	-220.9	-3.3	-213.8	3.8
INA:35TFMBA	"	-214.6	-221.3	-6.7	-239.3	-24.7
INA:MEFA	"	-261.9	-274.1	-12.1	-259.0	2.9
INA:FAMEE	"	-216.7	-184.1	32.6	-184.5	32.2
NA:FUA	-107.9	-228.0	-218.8	9.2	-218.0	10.0
NA:4HBA	"	-224.7	-231.0	-6.3	-216.7	7.9
NA:EPB	"	-216.3	-221.3	-5.0	-215.9	0.4
NA:2C4NBA	"	**-228.0**	**-226.8**	**1.3**	**-228.0**	**0.0**
NA:TOFA	"	-366.1	-379.5	-13.4	-363.2	2.9
NA:MEFA	"	**-359.8**	**-338.1**	**21.8**	**-348.9**	**10.9**
NA:NIFA	"	-262.3	-274.5	-12.1	-269.9	-7.5
NA:FURA	"	-318.0	-300.8	17.2	-318.8	-0.8
Validation set						
THP:SA	-140.2	-231.8	-236.4	-4.6	-251.9	-20.1
CAF:4F3NAN	-128.0	**-218.4**	**-221.3**	**-2.9**	**-220.9**	**-2.5**
INA:ADA	-114.2	-241.4	-220.5	20.9	**-187.4**	**54.0**
INA:GTA	"	-238.1	-225.5	12.6	-223.8	14.2
INA:4HBA	"	**-231.0**	**-208.8**	**22.2**	-206.7	24.3
NA:GTA	-107.9	-241.4	-220.5	20.9	-247.7	-6.3

A detailed examination of *T*m values of all 61 CCs are given in the Table 3. The best results (with minimum deviation from the actual *T*m values) for each MC in the training set is obtained for THP/DMA, CAF/SA_I, INA/2PPAR and NA:FURA, highlighted in green in the table. Large errors are observed for THP/PCA-II, CAF/2C5NAN, INA/SCA and NA/MEFA, highlighted in red. In the training set, an optimal low deviation of 0.1 K is obtained for THP and INA drug molecules with D-malic acid (DMA) and (R)-2-phenylpropionic acid (2PPAR) coformer respectively (Table 4). A large deviation of about 39.1 (K) is obtained for CAF, with 2-Chloro-5-nitroaniline (2C5NAN) coformer. On the other hand, with respect to the validation set, the smallest deviation is obtained for INA, with glutaric acid (GTA) coformer with ± 11.6 K deviation, whereas the biggest deviation is obtained for INA, with 4-hydroxybenzoic acid (4HBA) coformers (Table 3 & 4). Model–I depicts both positive and negative deviations from the experimental *T*m values. As an average, 3.61 (K) of positive deviation is obtained for THP CCs, 7.2 (K) of positive deviation is obtained for CAF CCs, –6.64 (K)

of negative deviation is obtained in INA CCs series, and 5.2 (K) positive deviation is obtained for NA CCs. The smallest deviation is obtained for THP, whereas the highest deviation is obtained for CAF CCs.

In the case of E_{latt} prediction of CCs, the best value for each MC in the training set is obtained for THP/GTA, CAF/GTA_I, INA/DLMDA and NA/2C4NBA, highlighted in green in the Table 5. The largest deviations are obtained for THP/GNA, CAF/4I3NAN, INA/4KPA and NA/MEFA, which are marked in red. The smallest deviation in the training set is obtained for THP, with glutaric acid (GTA) coformer having –0.4 kJ/mol deviation, while the largest deviation of 119.2 kJ/mol is obtained for 4–ketopimelic acid (4KPA) coformer in the INA series. On the other hand, the smallest deviation in the validation set is obtained for CAF/4F3NAN with –2.9 kJ/mol deviation, whereas the highest deviation is obtained for INA/4HBA with +22.2 kJ/mol error marked as green and red color respectively in the Table 5.

For the $CC_{CD,}$ the average relative deviation is 1.88% for the training set, and 0.99% for the validation set. The results are illustrated in the Figure 6, and the values are given in the Table 10. The best results in each MC series with respect to various coformers is obtained for THP/PCA–II, CAF/2F5NAN, INA/SCA and NA/2C4NBA, highlighted in green in Table 10. The biggest deviation is obtained for THP/DMA, CAF/4C3NAN, INA/3NBA, INA/CIA and NA/MEFA, marked by red. The best prediction without any deviation in the validation set is obtained for THP, with salicylic acid (SA) coformer; the highest deviation is obtained for INA, with 4-hydroxybenzoic acid coformer about –0.054 g/cm^3, highlighted in green and red respectively in Table 10.

Table 6. Sensitivity analysis of model–I. Mean squared error (MSE)/average relative error (ARE).

S.No	Removed input parameter	Training set error		Validation set error	
		MSE (%)	ARE (%)	MSE (%)	ARE (%)
1	API_{MW}	0.002	4.33	0.003	5.19
2	CF_{MW}	0.003	5.5	0.004	8.75
3	API_{BE}	0.002	4.49	0.002	6.35
4	CF_{BE}	0.002	4.41	0.002	4.86
5	CC_{BE}	0.002	4.47	0.003	7.74
6	API_{Tm}	0.002	4.50	0.002	5.64
7	CF_{Tm}	0.003	4.60	0.003	6.95
8	$CC_{\Delta pKa}$	0.002	4.40	0.002	5.90

A sensitivity analysis has been performed for model–I with respect to the input parameters. The importance of each input parameter has been investigated by performing the training and validation on seven input neurons by removing one at a time each of the input parameters. In doing so, eight individual models were constructed and the average relative deviation for the training set and the validation set respectively reported in Table 6. In each case, the performance of model–I is decreasing as illustrated by the increasing deviation values for both the training and validation sets. Accordingly, all eight input variables makes a valuable contribution to the prediction of *T*m, E_{latt} and CD of the CCs in model–I, and model–I satisfies the convergence criterion.

Model-II – Prediction of the cocrystal melting point Tm

Using the same eight input parameters as for model–I, the training process for model–II was stopped after reaching the convergence criterion, resulting in an average relative deviation of 1.93% (compare 2.49% obtained for model–I) for the training set and 2.23% (6.26% error obtained for model–I) for the validation set. The deviations are about the same for the validation and the training

set, and hence, there is no overfitting in the model. For individual values, in most cases model–II performs better than model–I. The CCs systems giving the lowest and highest deviation in the training set and the validation set for each MC for prediction of *T*m by model–II, are highlighted in green and red respectively in Table 3. The sensitivity analysis for model–II (Table 7) again reveal that all eight parameters are important for prediction of *T*m of CCs.

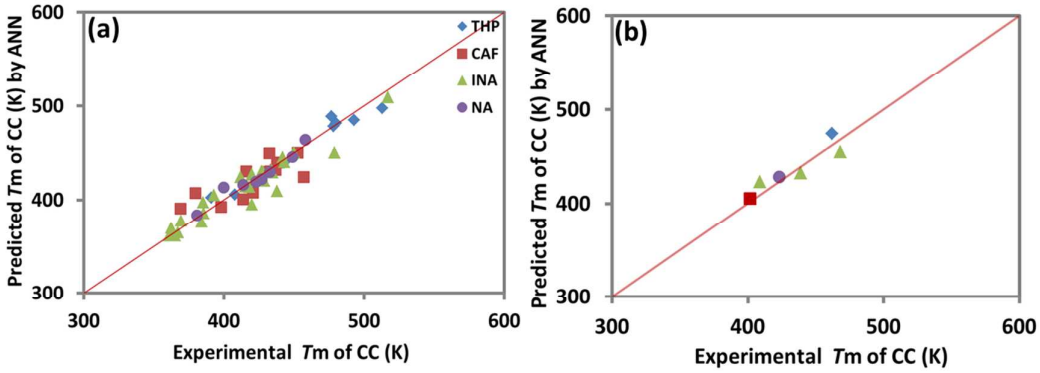


Figure 6. Model-II predicted *vs* experimental melting temperature (eigth input parameters): (a) training set, (b) validation set.

Table 7. Sensitivity analysis of model–II. Mean squared error (MSE)/average relative error (ARE).

S.No	Removed input parameter	Training set error		Validation set error	
		MSE (%)	**ARE** (%)	**MSE** (%)	**ARE** (%)
1	API_{MW}	0.001	2.61	0.002	3.63
2	CF_{MW}	0.000	2.10	0.004	6.47
3	API_{BE}	0.001	2.17	0.003	4.64
4	CF_{BE}	0.001	2.39	0.004	6.08
5	CC_{BE}	0.001	2.58	0.005	6.62
6	API_{Tm}	0.001	2.67	0.003	5.07
7	CF_{Tm}	0.001	2.73	0.002	4.90
8	$CC_{\Delta pKa}$	0.001	1.93	0.005	5.32

Model-III – Prediction of the cocrystal lattice energy (CC_{Elatt})

Using the same eight input neurons as for model–I, the training process for model–III stopped in a global minimum with 1.29% average relative deviation for the training set and 2.40% for the validation set (compare 6.21% and 4.58% respectively for model-I). The values for both the training and the validation sets are quite small, and have a small difference between them of about ~1.11%. Hence, model–III is well suited for prediction of E_{latt} of the CCs. In most cases, model–III performs better than model–I, the lowest and highest deviation are highlighted in green and red respectively in Table 5. The sensitivity analysis for model–III (Table 8), shows that all eight input parameters are important for the prediction of E_{latt} of the CCs.

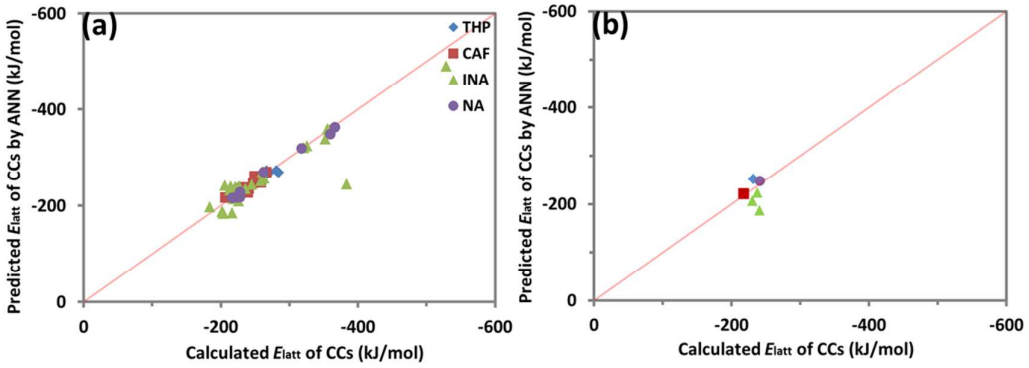


Figure 7. Model–III predicted *vs* experimental lattice energy (eight input parameters): (a) training set, (b) validation set.

Table 8. Sensitivity analysis of model–III. Mean squared error (MSE)/average relative error (ARE).

S.No	Removed input parameter	Training set error		Validation set error	
		MSE (%)	ARE (%)	MSE (%)	ARE (%)
1	MC_{MW}	0.002	3.88	0.002	9.17
2	CF_{MW}	0.008	9.25	0.014	18.83
3	MC_{BE}	0.001	4.42	0.009	17.02
4	CF_{BE}	0.001	3.77	0.003	10.54
5	CC_{BE}	0.002	4.08	0.003	9.08
6	MC_{Tm}	0.004	7.11	0.007	14.10
7	CF_{Tm}	0.005	5.72	0.001	6.77
8	$CC_{\Delta pKa}$	0.001	3.75	0.003	10.64

Model-IV – Prediction of the cocrystal crystal density (CC_{CD})

Using the same eight input neurons as for model–I, the traning process for model–IV was stopped at an average relative deviation of 1.03% for the training set and 1.77% validation set, respectively, as compare to 1.88%, and 0.99%, respectively for model–I. Remembering that a lower value for the validation set as compared to the training set is indicating that the process is not working well, it is concluded that model–IV is more appropriate for the prediction of the CD of the CCs. The best CCs systems with lowest deviation, and the CC systems with highest deviation are highlighted in green and red respectively, in Table 10. The sensitivity analysis of model–IV reveals that all eight parameters are important for the prediction of CD of the CCs.

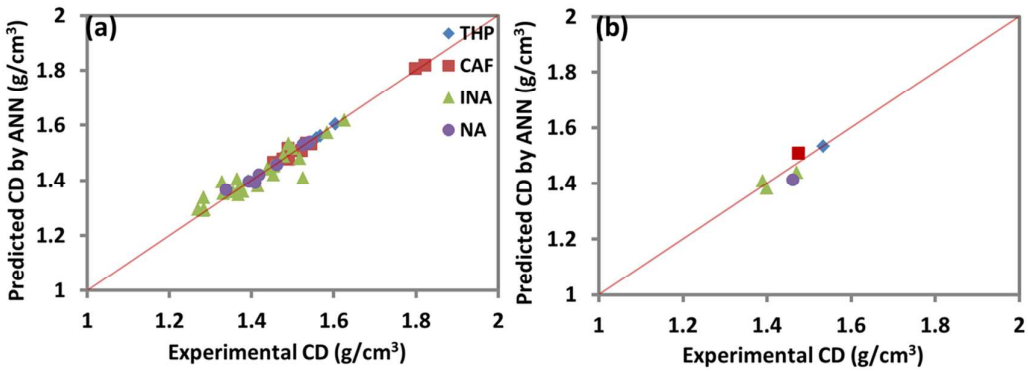


Figure 8. Model–IV predicted *vs* experimental crystal density (eigth input parameters): (a) training set, (b) validation set.

Table 9. Sensitivity analysis of model–IV. Mean squared error (MSE)/average relative error (ARE).

S.No	Removed input parameter	Training set error		Validation set error	
		MSE (%)	ARE (%)	MSE (%)	ARE (%)
1	MC_{MW}	0.000	1.2	0.000	1.98
2	CF_{MW}	0.001	2.71	0.001	3.33
3	MC_{BE}	0.000	1.15	0.000	2.11
4	CF_{BE}	0.000	1.55	0.001	2.52
5	CC_{BE}	0.000	1.29	0.000	1.84
6	MC_{Tm}	0.000	1.27	0.000	1.93
7	CF_{Tm}	0.000	1.69	0.001	2.99
8	$CC_{\Delta pKa}$	0.000	1.38	0.001	2.71

Table 10. Experimental and predicted crystal densities of the 61 CCs of CAF, THP, NA and INA model components.

Name of the CC	MC_{CD} g/cm^3	CC_{CD} (Exp) g/cm^3	CC_{CD} (Predicted) g/cm^3, model-I	CC_{CD} (pre) - CC_{CD} (exp) g/cm^3, model-I	CC_{CD} (Predicted) g/cm^3, model-IV	CC_{CD} (pre) - CC_{CD} (exp) g/cm^3, model-IV
Training set						
THP: DLMA	1.51	1.544	1.535	-0.009	1.538	-0.006
THP: DMA	"	**1.558**	**1.531**	**-0.027**	1.554	-0.004
THP:GTA	"	1.489	1.515	0.026	1.501	0.012
THP:GNA	"	1.567	1.557	-0.010	1.560	-0.007
THP:PCA-I	"	1.483	1.499	0.016	1.481	-0.002
THP:PCA-II	"	**1.491**	**1.488**	**-0.003**	**1.491**	**0.000**
THP:SAC	"	1.604	1.578	-0.026	1.603	-0.001
THP:URE	"	1.460	1.479	0.019	1.465	0.005
CAF:GTA_I	1.483	1.482	1.487	0.005	1.478	-0.004
CAF:GTA_II	"	1.486	1.498	0.012	1.481	-0.005
CAF: PCA	"	1.478	1.454	-0.024	**1.478**	**0.000**
CAF:4NAN	"	1.453	1.458	0.005	1.466	0.013
CAF:2I4NAN	"	1.822	1.791	-0.031	1.819	-0.003
CAF:2F5NAN	"	**1.489**	**1.489**	**0.000**	1.516	0.027
CAF:4C3NAN	"	**1.545**	**1.480**	**-0.065**	1.533	-0.012
CAF:2C5NAN	"	1.522	1.497	-0.025	1.508	-0.014
CAF:4I3NAN	"	1.799	1.761	-0.038	1.808	0.009
CAF:24DNBA	"	1.533	1.547	0.014	1.535	0.002
CAF:2F5NBA	"	1.489	1.487	-0.002	1.499	0.010
CAF:SA	"	1.490	1.482	-0.008	1.478	-0.012
CAF:SA_I	"	1.476	1.483	0.007	1.478	0.002
INA:OXA	1.347	1.584	1.553	-0.031	1.573	-0.011
INA:MLA	"	1.490	1.558	0.068	1.533	0.043
INA:SCA	"	**1.478**	**1.474**	**-0.004**	1.488	0.010
INA:PIA	"	1.331	1.344	0.013	1.352	0.021
INA:SUA	"	1.369	1.357	-0.012	1.370	0.001
INA:AZA	"	1.283	1.336	0.053	1.337	0.054
INA:FUA	"	1.500	1.473	-0.027	1.516	0.016
INA:4KPA	"	1.415	1.368	-0.047	1.379	-0.036
INA:12BDA	"	1.392	1.422	0.030	1.398	0.006
INA:SA	"	1.455	1.444	-0.011	1.454	-0.001
INA:3HBA	"	1.451	1.421	-0.030	1.450	-0.001
INA:4FBA	"	1.453	1.425	-0.028	1.420	-0.033
INA:3NBA	"	1.526	**1.410**	**-0.116**	**1.407**	**-0.119**
INA:2HEA	"	1.285	1.326	0.041	1.295	0.010
INA:CIA	"	1.365	**1.401**	**-0.116**	1.404	0.039
INA:CAA	"	1.517	1.463	-0.054	1.480	-0.037
INA:2PPARS	"	1.282	1.293	0.011	1.289	0.007
INA:2PPAR	"	1.270	1.296	0.026	1.294	0.024
INA:DLMDA	"	1.373	1.469	0.096	1.379	0.006
INA:CFA	"	1.356	1.338	-0.018	1.358	0.002
INA:REOL	"	1.373	1.389	0.016	1.367	-0.006

INA:HQ	"	1.378	1.386	0.008	1.359	-0.019
INA: 3NNDMABA	"	1.329	1.399	0.070	1.394	0.065
INA:35TFMBA	"	1.626	1.612	-0.014	1.619	-0.007
INA:MEFA	"	1.438	1.466	0.028	1.441	0.003
INA:FAMEE	"	1.368	1.308	-0.060	1.347	-0.021
NA:FUA	1.403	1.338	1.348	0.010	1.365	0.027
NA:4HBA	"	1.462	1.442	-0.020	1.458	-0.004
NA:EPB	"	1.418	1.456	0.038	1.421	0.003
NA:2C4NBA	"	**1.393**	**1.391**	**-0.002**	1.397	0.004
NA:TOFA	"	1.526	1.507	-0.019	1.530	0.004
NA:MEFA	"	**1.409**	**1.365**	**-0.044**	1.392	-0.017
NA:NIFA	"	1.533	1.561	0.028	1.535	0.002
NA:FURA	"	1.540	1.514	-0.026	1.539	-0.001
Validation set						
THP:SA	1.51	1.534	**1.534**	**0.000**	**1.533**	**-0.001**
CAF:4F3NAN	1.483	1.474	1.470	-0.004	1.508	0.034
INA:ADA	1.347	1.390	1.404	0.014	1.410	0.020
INA:GTA	"	1.399	1.398	-0.001	1.382	-0.017
INA:4HBA	"	1.471	**1.417**	**-0.054**	1.439	-0.032
NA:GTA	1.403	1.461	1.467	0.006	**1.413**	**-0.048**

Conclusions

In this study, a machine learning Artificial Neural Network (ANN) approach has been used to create four different models for prediction of solid state properties of cocrystals; *i.e.* melting point, lattice energy and crystal density. Notably the models are not using input information that require manufacturing of the cocrystal, but only data for the pure compounds and a simulated 1:1 MC–coformer binding energy, altogether 8 different input parameters. By model–II, model–III and model–IV each respective output variable, *i.e.* melting temperature, lattice enery and crystal density can be predicted to an average relative deviation of 1.93%, 1.29% and 1.03% respectively for the training sets, and 2.23%, 2.40% and 1.77% respectively for the validation sets. By model–I all three output variables are predicted by one and the same model however to an overall lower accuracy. The average relative deviation is 2.49%, 6.21% and 1.88% respectively for the melting temperature, lattice energy and crystal density training set, and correspondingly 6.26%, 4.58% and 0.99% respectively for the validation set.

Author information

Corresponding Author

*E-mail: Ake.Rasmuson@ul.ie

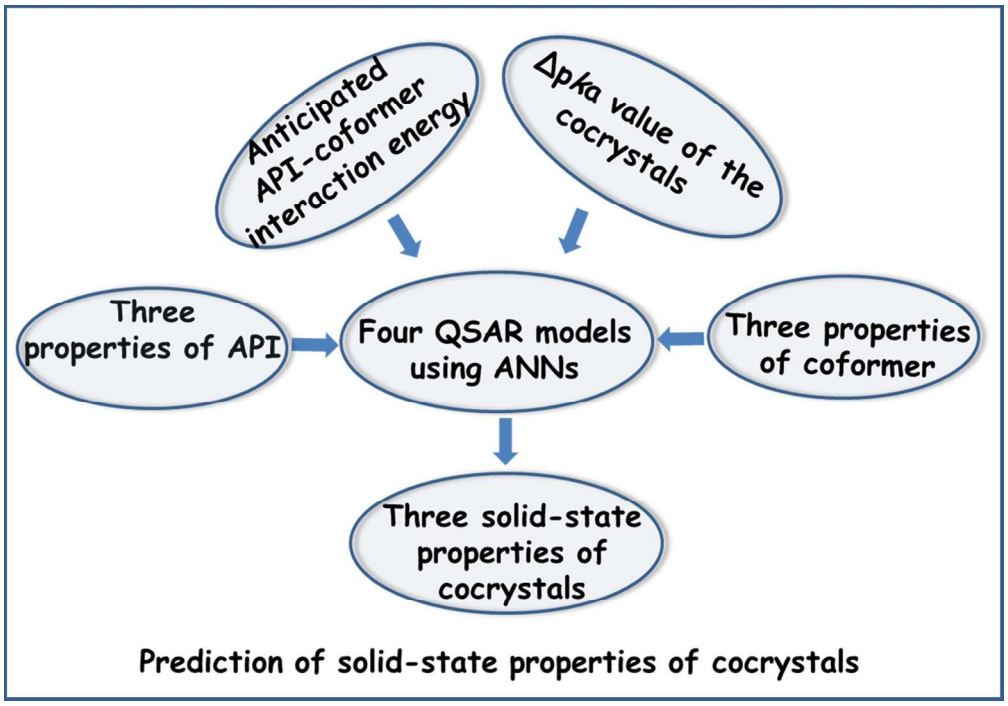
Acknowledgment

The authors acknowledge financial support from the Science Foundation Ireland, Grant number: 12/RC/2275.

References

1. Bergström, C. A. S.; Norinder, U.; Luthman, K.; Artursson, P. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1177–1185.
2. Batisai, E.; Ayamine, A.; Kilinkissa, O. E. Y.; Báthori, N. B. *CrystEngComm*. **2014**, *16*, 9992–9998.

3. Mao, F.; Kong, Q.; Ni, W.; Xu, X.; Ling, D.; Lu, Z.; Li, J. *ChemistryOpen*. **2016**, *5*, 357 – 368.
4. Compernolle, S.; Ceulemans, K.; Müller, J.–F. *Atmos. Chem. Phys*. **2011**, *11*, 9431–9450.
5. Simamora, P.; Miller, A. H.; Yalkowsky, S. H. *J. Chem. Inf Comput. Sci*. **1993**, *33*, 437–444.
6. Chu, K. A.; Yalkowsky, S. H. *Int. J. Pharm*, **2009**, *373*, 24–40.
7. Le, T.; Epa, V. C.; Burden, F. R.; Winkler, D. A. *Chem. Rev*. **2012**, *112*, 2889–2919.
8. Tetko, I. V.; Sushko, Y.; Novotarskyi, S.; Patiny, L.; Kondratov, I.; Petrenko, A. E.; Charochkina, L.; Asiri, A. M. *J. Chem. Inf. Model*. **2014**, *54*, 3320−3329.
9. Katritzky, A. R.; Jain, R.; Lomaka, A.; Petrukhin, R.; Maran, U.; Karelson, M. *Cryst. Growth Des*. **2001**, *1*, 261–265.
10. Modarresi, H.; Dearden, J. C.; Modarress, H. *J. Chem. Inf. Model*. **2006**, *46*, 930–936.
11. Bhat, A. U.; Merchant, S. S.; Bhagwat, S. S. *Ind. Eng. Chem. Res*. **2008**, *47*, 920–925.
12. Aakeröy, C. B.; Panikkattu, S.; DeHaven, B.; Desper, J. *CrystEngComm*, **2013**, *15*, 463–470.
13. Aakeröy, C. B.; Forbes, S.; Desper, J. *CrystEngComm.* **2014**, *16*, 5870–5877.
14. Bond, A. D. *CrystEngComm*. **2006**, *8*, 333–337.
15. Cholakov, G.; Stateva, R. P.; Shacham, M.; Brauner, N. *AIChE J*. **2007**, *53*, 150–159.
16. Sanphui, P.; Devi, V. K.; Clara, D.; Malviya, N.; Ganguly, S.; Desiraju, G. R. *Mol. Pharm*. **2015**, *12*, 1615–1622.
17. Cherukuvada, S.; Jagadeesh babu, N.; Nangia, A. *J. Pharm. Sci*. **2011**, *100*, 3233–3244.
18. Krishna, G. R.; Shi, L.; Bag, P. P.; Sun, C. C.; Reddy, C. M. *Cryst. Growth Des*. **2015**, *15*, 1827–1832.
19. Duggirala, N. K.; Perry, M. L.; Almarsson, ö.; Zaworotko. M. J. *Chem. Commun*. **2016**, *52*, 640–655.
20. Aakeröy, C. B.; Forbes, S.; Desper, J. *J. Am. Chem. Soc*. **2009**, *131*, 17048–17049.
21. Perlovich, G. L. *CrystEngComm.* **2015**, *17*, 7019–7028.
22. Gamidi, R. K.; Rasmuson, Å. C. *Cryst. Growth Des*. **2017**, *17*, 175–182.
23. Agatonovic-Kustrin, S.; Beresford, R. *J. Pharm. Biomed. Anal*. **2000**, *22*, 717–727.
24. Kumar. K. V.; Martins, P.; Rocha, F. *Ind. Eng. Chem. Res*. **2008**, *47*, 4917–4923.
25. Kumar. K. V. *Ind. Eng. Chem. Res*. **2009**, *48*, 4160–4164.
26. Kumar. V. K.; Porkodi, K.; Avila Rondon, R. L.; Rocha, F. *Ind. Eng. Chem. Res*. **2008**, *47*, 486–490.
27. Kůrková, V. *Neural Netw.* **1992**, *5*, 501–506.
28. Saha, A. Neural Network Models in Excel for Prediction and Classification; https://www.sites.google.com/site/sayhello2angshu/dminexcel.
29. Dubey, R.; Pavan, M. S.; Guru Row, T. N.; Desiraju, G. R. *IUCrJ*, **2014**, *1*, 8–18.
30. Feng, R–z.; Zhang, S–h.; Ren, F–d.; Gou, R–j.; Gao, L. *J. Mol. Model*, **2016**, *22,* 123–136.
31. Desiraju, G. R. *Angew. Chem. Int. Ed*. **1995**, *34*, 2311–2327.
32. Shattock, T. R.; Arora, K. K.; Vishweshwar, P.; Zaworotko, M. J. *Cryst. Growth Des*. **2008**, 8, 4533–4545.
33. Bolla, G.; Nangia, A. *Chem. Commun*. **2015**, *51*, 15578–15581.



Prediction of Solid State Properties of Co-crystals

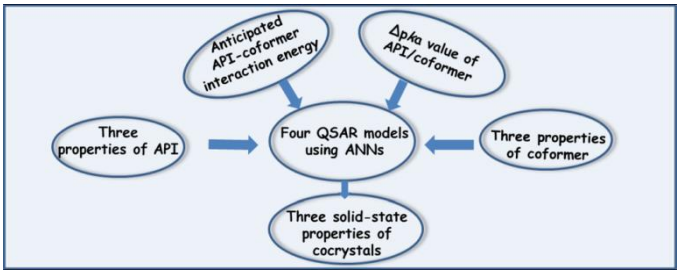
213x147mm (150 x 150 DPI)

For Table of Contents Use Only

Prediction of Solid State Properties of Co-crystals using Artificial Neural Network Modelling

Gamidi Rama Krishna, Marko Ukrainczyk, Jacek Zeglinski and Åke. C. Rasmuson*

Department of Chemical and Environmental Science, Synthesis and Solid State Pharmaceutical Centre, Bernal Institute, University of Limerick, Limerick, Ireland



Synopsis: Artificial Neural Network (ANNs) models have been developed for the prediction of solid-state properties of cocrystals. The models use in total eight input parameters: three input parameters for each of the API and the conformer, the p*K*a difference between the API and the conformer, and the energy of the anticipated 1:1 API–conformer binding as calculated by a force field method.